VIEWPOINT DISTORTION COMPENSATION IN PRACTICAL SURVEILLANCE SYSTEMS

Ognjen Arandjelović, Duc-Son Pham, and Svetha Venkatesh

Centre for Pattern Recognition and Data Analytics, Deakin University, Australia

ABSTRACT

Our aim is to estimate the perspective-effected geometric distortion of a scene from a video feed. In contrast to all previous work we wish to achieve this using from low-level, spatio-temporally local motion features used in commercial semi-automatic surveillance systems. We: (i) describe a dense algorithm which uses motion features to estimate the perspective distortion at each image locus and then polls all such local estimates to arrive at the globally best estimate, (ii) present an alternative coarse algorithm which subdivides the image frame into blocks, and uses motion features to derive block-specific motion characteristics and constrain the relationships between these characteristics, with the perspective estimate emerging as a result of a global optimization scheme, and (iii) report the results of an evaluation using nine large sets acquired using existing close-circuit television (CCTV) cameras. Our findings demonstrate that both of the proposed methods are successful, their accuracy matching that of human labelling using complete visual data.

Index Terms— Surveillance, novelty, normalization.

1. INTRODUCTION

In this paper we address the problem of inferring the perspective of a static scene with moving objects. In its general form, this problem has been extensively studied in the past. However, unlike in previous work, here our aim is to accomplish this under a set of constraints which have not been considered previously in the literature. Specifically, we wish to estimate the dominant perspective of a scene using low-level motion features, local both in spatial and temporal sense.

Our motivation for this problem and the source of the outline constraints stems from the operational requirements of many semi-automatic CCTV-based surveillance systems [1, 2]. In particular, we are interested in systems which occupy the largest portion of the commercial market today, and which detect abnormalities in video feeds on a semantically low-level (i.e. without 'understanding' or interpreting the nature of the detected abnormality). They accomplish this by extracting low-level features from which a scene-specific statistical model of normal motion in a scene is learnt. Spatio-temporally local motion is universally used as the basis for the elementary features over which learning is performed. These features have been proven successful in practice and they fit the computational constraints imposed by the need for most of the processing and data storage to be done locally using the camera's on-board hardware. This particular constraint emerges as a consequence of the large scale of many practical CCTV systems which for scalability reasons have a distributed architecture with minimal reliance on communication with the central hub.

As detailed in the next section, none of the previous work has addressed the problem of perspective estimation using the setup described above. In this paper we introduce two algorithms that solve the problem using two different approaches.

2. RELATED WORK

The problem addressed in the present paper is intimately related with the corpus of work on estimating the position and orientation of a camera and the recovery of 3D scene structure. This has been an active topic of research since the early days of computer vision, resulting in a number of mature techniques which are now widely used in practice (in film production, for example).

When there are known correspondences between 3D world points and their 2D projections, there is a series of algorithms that have been described in the literature which successfully handle cases for different numbers of available correspondences using different (usually iterative) optimization schemes [3, 4]. When no explicit world-to-camera mapping data is available but the camera is in motion and the scene mostly static, perceived (image) motion and different types of constraints (e.g. probabilistic, epipolar, or motion parallax based) can be used instead [5, 6]. Stronger yet constraints must be employed when it is not possible to obtain 3D-to-2D correspondences and the camera is static. For example for built-up scenes outline maps of buildings have been used with success by a number of researchers [7]. Similarly, in urban or indoor scenes the presence of many parallel lines (e.g corridor or street boundaries) and their convergence towards the same vanishing point can be used to estimate the perspective [8, 9]. Yet others learn appearance of different types of elementary structures which allows them to build an approximate model of the scene [10, 11].

As we shall see in the next section, none of these approaches meet all the requirements of a practical CCTV system. The cameras we are dealing with are static, their large

number makes an elaborate setup procedure required to obtain 3D-to-2D point correspondences impractical, and the algorithm deployed must be sufficiently robust to handle a wide variety of scenes and poor image/video quality thus prohibiting the reliance on the presence of strong cues such as parallel lines or specific object types.

3. BACKGROUND – OPERATIONAL CONSTRAINTS

Computer-assisted video surveillance data analysis is of major commercial and law enforcement interest. Unsurprisingly, this interest has spurred a major research effort as well; topics such as human detection [12] and tracking [13], crowd analysis [14], activity recognition [15], and others, have been attracting a significant amount of attention. Nonetheless, commercially available systems are still in relative infancy with most of the methods described in the academic literature still not sufficiently mature for practical deployment. Achieving reliable performance for scenes of different types, a wide range of foreground object classes which themselves exhibit major within-class variability, changing illumination conditions and shadows, and the breadth of possible events of interest all continue to pose a major challenge.

On a broad scale, systems currently available on the market can be grouped into two categories in terms of their approach. The first group focuses on a relatively small, predefined and well understood subset of events or behaviours of interest [16, 17]. Examples include the detection of unattended baggage, violent behaviour, or specific incidences of vandalism. While suitable in certain circumstances, the narrow focus of these systems prohibits their applicability in less constrained environments in which a more general capability is required. In addition, these approaches tend to be computationally expensive and error prone, often requiring fine tuning by skilled technicians. This is not practical in many circumstances, for example when hundreds of cameras need to be deployed as is often the case with CCTV systems operated by municipal authorities. The second group of systems approaches the problem of detecting suspicious events at a semantically lower level [18, 19, 20, 21]. Their central paradigm is that an unusual behaviour at a high semantic level will be associated with statistically unusual patterns (also 'behaviour' in a sense) at a low semantic level – the level of elementary image/video features. Thus methods of this group detect events of interest by learning the scope of normal variability of low-level patterns and alerting to anything that does not conform to this model of what is expected in a scene, without 'understanding' or interpreting the nature of the event itself. These methods nearly universally employ motion features [22], rather than appearance features (such as SIFT [14], SURF [23] or GLOH [24]). This is understandable considering that novelties of interest – events and behaviours – are inherently associated with motion, whereas appearance in a scene can vary greatly without the change being associated with anything of potential interest. For example, diurnal or seasonal changes in illumination can create major appearance differences while, on the other hand, resulting in no motion features due to their low temporal frequency; cars (or similarly pedestrians) passing down the road differ greatly one from another, yet it is reasonable to expect a much more constrained range of variability of their motion if they obey the rules of traffic.

In this paper we focus on the methods of the second group described above. We are particularly motivated to do so by their representation on the market on the one hand, and the lower amount of research attention they have attracted in comparison with the first group, on the other.

3.1. Available data – description and constraints

The surveillance analysis methods we are interested in all start with the same procedure for feature extraction. As video data is acquired, firstly a dense optical flow field is computed. Then, to reduce the amount of data that needs to be processed, stored, or transmitted, a thresholding operation is performed. This results in a sparse optical flow field whereby only those flow vectors whose magnitude exceeds a certain value are retained; non-maximum suppression is applied here as well. Normal variability within a scene and subsequent novelty detection are achieved using this data.

One problem with the data pre-processing procedure described above is that it does not consider the effects of perspective distortion on the appearance of the scene. Most obviously, the thresholding applied on the optical flow vectors should be dependent on their location in the image plane. At present, the threshold is set sufficiently low to pick up potential motions of interest in the most distant parts of the scene, which has the disadvantage of unnecessarily increasing the amount of data extracted in the regions closer to the camera. Equally, any subsequent learning could be made more robust if it took the effects of perspective into account. Our goal in this paper is to empower the existing surveillance analytics with this information, that is, to estimate the dominant perspective scaling effect in a scene using the described sparse optical flow field.

In other words, our input data comprises a sequence of sets of motion vectors. Two temporally-neighbouring sets are separated by a uniform time interval Δt, which is governed by the frame rate of the camera (5–15 fps). Motion vector sets are in general of different sizes, and each set describes the sparse optical flow field at a particular time instant. All motion vectors are also spatially localized in the image plane.

4. PROPOSED METHODS

In the previous sections we explained why none of the existing methodologies for perspective estimation are adequate for the deployment in the operational setting of many CCTV systems. In this section we introduce two novel methods that

solve the problem at hand effectively. The two solutions we propose can be contrasted by the spatial scale at which inference is made. The first algorithm adopts a dense approach, whereby a perspective estimate is made at all image plane locations and the final estimate for the whole scene emerges through a consensus of these estimates. The second algorithm operates at a larger scale. It divides the image plane into blocks and performs inference by accounting for motion statistics in different image blocks and the constraints between neighbouring blocks. At the bottom-most level, both methods are based on the observation that the motion of an object farther in a scene exhibits itself as a linearly scaled version of the apparent motion which would be observed if the object was closer to the camera. The key challenge arises from the fact that in practice it is impossible to perform such controlled calibration. Instead we formulate sets of constraints, of a different type for the dense algorithm and the block algorithm, which allow us to infer the perspective-induced scaling from the projected motion observed in the image plane.

4.1. Dense approach – pixel level

The first method we describe is dense in the sense that a local perspective estimation is made at every image locus (on the scale of a pixel). These local estimates are then polled and the overall estimate of the perspective distortion made through the consensus across the image plane. It is important to observe the assumption which is needed to make this approach sound. Specifically, we assume that in terms of its area when projected onto the image plane the scene is dominated by the ground plane in which motion takes place. This is a sensible assumption to make considering the purpose of CCTV cameras and the strategic manner of their placement.

We adopt the standard non-skew pin-hole camera model. Without loss of generality, if the origin of a right-hand coordinate system is placed at the focal point and the $x-y$ plane made parallel to the image plane at the focal length f, a 3D point $[x, y, z]^T$ is projected to the image plane point $[u, v]^T$ as follows:

$$u = f \cdot k_u \cdot \frac{x}{z} \qquad v = f \cdot k_v \cdot \frac{y}{z}, \tag{1}$$

where k_u and k_v are the camera's horizontal and vertical scaling parameters. The internal camera parameters – namely its focal length f, and the scaling parameters k_u and k_v – are of no interest to us since the normalization we seek is camera specific and is applied on a camera-by-camera basis.

Our goal is to estimate the dominant perspective-effected change in scale in the image plane. We formalize this by saying that we seek to estimate the quantity ζ we defined as:

$$\zeta = \frac{dz/z}{dv}. \tag{2}$$

In other words, we wish to know how the distance of an object in the scene (or, equivalently, its perceived scale) changes with the change in its projected image locus. Here we show that ζ can be estimated from spatially and temporally local pure motion features (described in the previous section) only.

Consider the change in the vertical component of the projected (image plane) velocity of a point which corresponds to a world point moving in the ground plane. It is proportional to the component of the point's world velocity in the direction of the camera $\frac{dw}{dt}$ (w is the projection of z onto the ground plane) and inversely proportional to its distance z from the camera's focal point. In other words:

$$d\dot{v} = c \cdot \frac{dw}{dt} \cdot \left[\frac{1}{z+dz} - \frac{1}{z}\right] = c \cdot \frac{dw}{dt} \cdot \frac{dz}{z^2}, \tag{3}$$

where we introduce the constant c to 'bundle' different constants for the sake of reducing clutter. By substitution in (2) we can derive:

$$\zeta = \frac{dz/z}{dv} = \frac{z\, d\dot{v}}{c \cdot \frac{dw}{dt} \cdot dv}. \tag{4}$$

Similarly, since:

$$\dot{v} = \frac{dv}{dt} = c \cdot \frac{dw/dt}{z}, \tag{5}$$

the expression in (4) can be further simplified to:

$$\zeta = \frac{dz/z}{dv} = \frac{d\dot{v}}{dv} \times \frac{1}{\dot{v}} \tag{6}$$

Succinctly expressed, this result shows that in the context under consideration, perspective-caused scaling in the image plane can be estimated using the rate of change of velocity of a feature in uniform world motion across the image, normalized by its observed, image plane velocity.

4.1.1. Model outliers

Throughout our derivation in the preceding section it was assumed that the observed world motion was piecewise uniform. While it can be reasonably expected that this assumption will be satisfied in most cases, it is equally true that it will undoubtedly often be invalidated as well. For example, a walking pedestrian may increase the speed of walking upon spotting a bus, or stop because something has attracted his/her attention. However, these are not systemic outliers – given sufficient data, it can be expected that opposite direction (acceleration vs. deceleration) deviations from the uniform motion will cancel each other out. A more serious challenge is posed by systemic outliers which emerge as a consequence of structural aspects inherent in the scene. For example, a corner in the road will consistently invalidate the assumption of uniform motion as cars entering the bend will slow down. Fortunately, this challenge is readily addressed in our framework considering that the dense nature of our method offers a large pool of quasi-independent perspective estimates. In particular, when polling different image locations, we aggregate all estimates in a vector, sort them and reject the top and bottom 15% quantiles before computing the mean of the remainder to arrive at the global estimate.

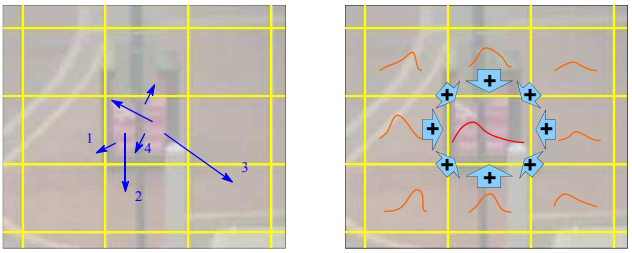
4.2. Coarse approach – block level

While successful in practice, as we will show in the next section, intuitively speaking the scale at which the method proposed in the previous section operates seems excessively fine considering the global nature of the assumption of one dominant perspective effect. This is further supported by the observation that in principle the same quantity ζ is estimated at every location in the image plane. Yet, in the proposed framework this polling is necessary as a means of rejecting outlier loci which do not satisfy the assumption of (on average) uniform world velocity of the corresponding features.

A simple way of reducing the number of image plane locations considered, and with it the computational cost of the method, could be implemented by prioritizing certain locations over others. For example, locations at which the amount of motion observed falls below a specified threshold could be reasonably considered as less reliable 'voters' and thus rejected from being polled at all. However this approach still does not address the fundamental overkill which working on the pixel level appears to be.

The method we describe next uses a division of the image plane into equally-sized rectangular blocks. This is illustrated in Figure 1(a). Each block is treated as a unit, that is, any property pertaining to the block is associated with the block as a whole. The key idea underlying our approach is that motion statistics in different blocks can be related to one another. Most notably this is true for neighbouring blocks. We exploit this observation by recording nine measurements for each block. The first of these, $m_{i,j}$, is the average velocity magnitude within the block (i and j are the vertical and horizontal indices of the block). For example, in Figure 1(a) all of the shown motion vectors would contribute to the central block's mean motion magnitude; notice that we compute the mean of all magnitudes, rather than the magnitude of the mean motion vector. The remaining eight *transition* measurements associated with the block, $\rho_{i,j}^{(\dots)}$, quantify the relatedness of the motion within the block and the motions in the block's 8-neighbourhood. A specific $\rho_{i,j}^{(\dots)}$, say $\rho_{i,j}^{tl}$, indicates the proportion of motion vectors which end in the block under the consideration and originate in its top-left neighbour ('tl'; similar indexing is used to denote the remaining directions, namely right and bottom, and the possible combinations). For example, in Figure 1(a) the motion vector labelled '1' contributes to $\rho_{i,j-1}^{(r)}$, the vector labelled '2' to $\rho_{i+1,j}^{(t)}$, the vector labelled '3' to $\rho_{i,j+1}^{(l)}$ and $\rho_{i+1,j+1}^{(tl)}$, and the vector labelled '4' to no transition measurement at all (but it does contribute to $m_{i,j}$, of course).

Our idea is to formulate the problem of estimating the relative scaling between neighbouring rows of blocks in the form of an optimization task. Specifically, for the block indexed by (i,j) we can write:



(a) (b)

Fig. 1. Conceptual illustration of the proposed coarse algorithm. The key idea is to count motion vectors resulting in transitions between blocks into which the image is divided (a), and use thus estimate transition statistics together with the motion statistics within each block to formulate perspective scaling constraints between different blocks (b).

$$\begin{aligned} m_{i,j} =& \rho_{i,j}^{(l)}\, m_{i,j-1} + \rho_{i,j}^{(r)}\, m_{i,j+1} + \\ & \left[\rho_{i,j}^{(tl)}\, m_{i-1,j} + \rho_{i,j}^{(t)}\, m_{i-1,j} + \rho_{i,j}^{(tr)}\, m_{i-1,j}\right] \times \omega + \\ & \left[\rho_{i,j}^{(bl)}\, m_{i+1,j} + \rho_{i,j}^{(b)}\, m_{i+1,j} + \rho_{i,j}^{(br)} m_{i+1,j}\right] \times \omega^{-1}, \end{aligned} \quad (7)$$

where ω is the factor quantifying perspective-effected scale change between consecutive rows of blocks. It is simple to show that it is related to the previously introduced ζ as:

$$\zeta = \frac{1-\omega^{-1/2}}{h}, \quad (8)$$

where h is the height of a block in pixels.

What (7) is saying is that the motion observed within (i,j) is a mixture of motions in the neighbouring blocks, weighted by the proportional contribution of each block (via different $\rho_{i,j}^{(\dots)}$) and where applicable the perspective-induced scaling ω. This is illustrated conceptually in Figure 1(b).

Since a constraint in the form of (7) can be written for each non-boundary block, thereby resulting in $(n-1)\times(m-1)$ equations with only a single unknown ω, the system is over-determined. Thus we seek the solution which gives the least L_2-norm error for the error vector comprising the differences between the left and right-hand sides of (7) for different blocks. The optimal solution is readily obtained by employing one of a number of standard iterative schemes. Nevertheless, we found that a comparably accurate result could be obtained using an approximation which can be computed in the closed form. In particular we exploit the observation that the viewing setup of CCTV cameras is such that the scaling factor between neighbouring block rows ω will be close to 1. Thus writing ω as $\omega = 1 + \Delta\omega$ and using the first-order Taylor expansion whereby ω^{-1} can be approximated as $\omega^{-1} \approx 1 - \Delta\omega$

allows for (7) to be replaced by its approximate form:

$$\begin{aligned} m_{i,j} \approx & \rho_{i,j}^{(l)}\, m_{i,j-1} + \rho_{i,j}^{(r)}\, m_{i,j+1} + \rho_{i,j}^{(tl)}\, m_{i-1,j} + \\ & \rho_{i,j}^{(t)}\, m_{i-1,j} + \rho_{i,j}^{(tr)}\, m_{i-1,j} + \rho_{i,j}^{(bl)}\, m_{i+1,j} + \\ & \rho_{i,j}^{(b)}\, m_{i+1,j} + \rho_{i,j}^{(br)} m_{i+1,j} + \\ & \Big[\rho_{i,j}^{(tl)}\, m_{i-1,j} + \rho_{i,j}^{(t)}\, m_{i-1,j} + \rho_{i,j}^{(tr)}\, m_{i-1,j} - \\ & \rho_{i,j}^{(bl)}\, m_{i+1,j} + \rho_{i,j}^{(b)}\, m_{i+1,j} + \rho_{i,j}^{(br)} m_{i+1,j}\Big] \times \Delta\omega, \end{aligned} \quad (9)$$

Clearly this is now a set of linear equations which is readily solved in the closed form for $\Delta\omega$ which minimizes the L_2-norm error using the corresponding pseudo-inverse matrix.

5. EVALUATION AND RESULTS

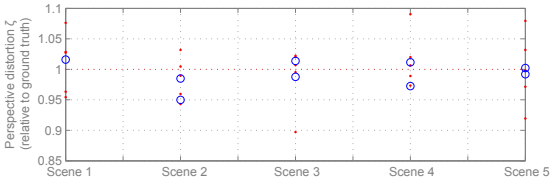
To assess the effectiveness of the proposed algorithms, we evaluated their performance on nine large 'real-world' data sets. It is important to emphasize that the data we used was not acquired for the purpose of the present work nor were the cameras installed with the same intention. Rather, we used data which was acquired using existing, operational surveillance systems. In particular, our data comes from five operational CCTV cameras in three major cities. Table 1 provides a summary the key statistics of the nine data sets, all of which were produced from original video feeds in 352×288 pixel resolution using the threshold of 1.5 pixels to arrive at a sparse optic flow field from the initial dense computation using the algorithm of Lucas and Kanade [25]. For our coarse, block-based method we used a grid of 10×10 blocks.

Table 1. Key statistics of the nine real-world data sets used in our evaluation. These were acquired from five operational CCTV cameras.

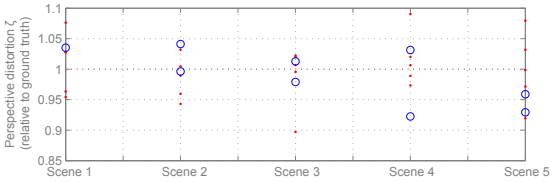
Scene	Data set	Duration of data acquisition	Frame rate	Avg. features per frame
Scene 1	1	10 min	5 fps	40.0
Scene 2	1	2 h	15 fps	18.1
	2	2 h	15 fps	14.1
Scene 3	1	2 h	15 fps	158.8
	2	2 h	15 fps	209.8
Scene 4	1	2 h	15 fps	141.0
	2	2 h	15 fps	161.1
Scene 5	1	2 h	15 fps	394.9
	2	2 h	15 fps	585.0

Considering that we could not obtain ground truth data directly, we used quasi-ground truth estimated using manually localized key points in the image plane. This was achieved by marking world-equidistant points at different distances from the camera. To increase the accuracy of thus obtained perspective estimates as well as to quantify the reliability of labelling, we asked for the labelling to be done by five different people for each scene.

We started our evaluation by looking at the overall performance of the two proposed algorithms when applied to each of the nine data sets. A summary of the key results is shown in Figure 2(a) and Figure 2(b) for the dense and coarse model based algorithms respectively. Each small red dot corresponds to the estimate of the perspective distortion coefficient ζ estimated from the input of one human labeller of quasi-ground truth; thus there are five red dots for each scene, as we used five labellers. The blue circles represent the estimates produced by the proposed algorithm. There are two blue circles per scene for all scenes except for the first one ('Scene 1') as for all but this scene we had two data sets of motion vectors. Note that the ordinate values have been normalized so that the values of ζ for different scenes could be visualized on the same graph. Specifically, we applied scaling to the displayed values which makes the mean value of the estimate of ζ resulting from human labelling equal to unity. The first thing to observe from Figures 2(a) and 2(b) is the outstanding performance of both of our algorithms. This is witnessed by the proximity of all automatically produced estimates of ζ to unity which is well within the range of deviation of human-based estimates. Secondly, the accuracy of our algorithms, as well as the soundness of the premises underlying their derivation is further corroborated by their mutual agreement. Again, the difference in their output is smaller than the difference of estimates given by different humans. Quantitative analysis suggests that the performance of the dense approach somewhat exceeds that of the coarse alternative (by less than 5%). While this is perhaps to be expected, considering the finer grained nature of the method and the robustness achieved by polling a large number of quasi-independent estimates, the difference is not significant.



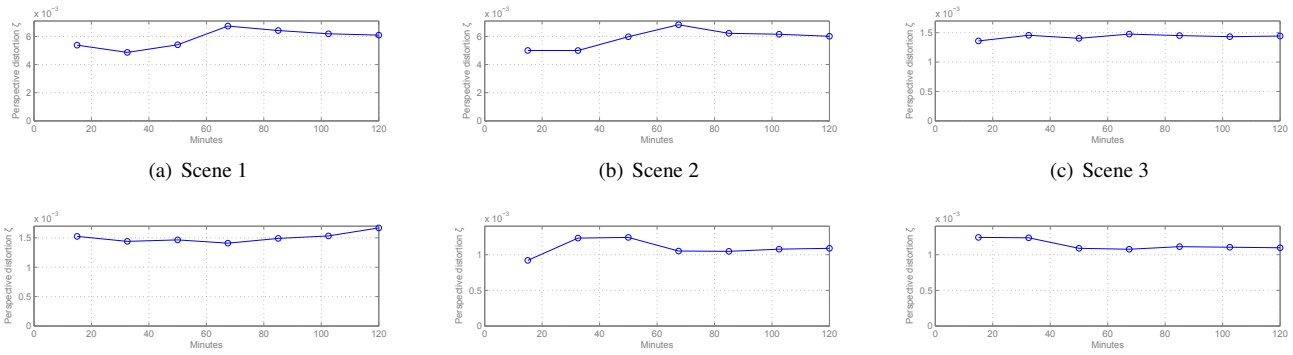
(a) Dense algorithm



(b) Coarse algorithm

Fig. 2. Performance of the two proposed algorithms. Red dots are human estimates of ζ; blue circles are estimates by the proposed algorithm. Ordinate values are normalized so that the values of ζ for different scenes could be visualized on the same graph – we applied scaling which makes the mean value of the human estimate of ζ equal to unity.

In the next set of experiments we were interested in exam-



(a) Scene 1 (b) Scene 2 (c) Scene 3

(d) Scene 4 (e) Scene 5 (f) Scene 6

Fig. 3. Variation of the scaling coefficient ζ with time i.e. with the accumulation of motion data.

ining how much data is needed to reach a reliable estimate of ζ. To do this, using each data set we produced seven different estimates – the first one after using only the first 12.5% of the data, the second one after using the first 25% of the data, etc. The results are shown in Figure 3 (only the results obtained using the dense model-based algorithm are shown as the results of the coarse algorithm were so similar that they could not be displayed effectively on the same plots). As expected, the greatest inaccuracy, as well as the greatest change with newly acquired data, is exhibited at the beginning, when the least amount of data is available for inference. Remarkably, in all cases, the convergence towards the correct value is fast with the error lower than 5% achieved within 90 minutes of data acquisition and in most cases in half of that time. For further analysis see [26].

6. REFERENCES

[1] D. Pham, O. Arandjelović, and S. Venkatesh, "Detection of dynamic background due to swaying movements from motion features.," *TIP*, 2015.

[2] O. Arandjelović, D. Pham, and S. Venkatesh, "Two maximum entropy based algorithms for running quantile estimation in non-stationary data streams.," *TCSVT*, 2015.

[3] F. Kahl and D. Henrion, "Globally optimal estimates for geometric reconstruction problems.," *IJCV*, 2007.

[4] J. A. Hesch and S. I. Roumeliotis, "A direct least-squares (DLS) method for PnP.," *ICCV*, 2011.

[5] X. Armangu, H. Arajo, and J. Salvi, "A review on egomotion by means of differential epipolar geometry applied to the movement of a mobile robot.," *PR*, 2003.

[6] J. Domke and Y. Aloimonos, "A probabilistic framework for correspondence and egomotion.," *ICCVW*, 2005.

[7] T.-J. Cham, A. Ciptadi, W.-C. Tan, M.-T. Pham, and L.-T. Chia, "Estimating camera pose from a single urban ground-view omnidirectional image and a 2D building outline map.," *CVPR*, 2010.

[8] J.-C. Bazin, Y. Seo, C. Demonceaux, P. Vasseur, K. Ikeuchi, I. Kweon, and M. Pollefeys, "Globally optimal line clustering and vanishing point estimation in Manhattan world.," *CVPR*, 2012.

[9] A. Herout, I. Szentandrási, M. Zachariás, M. Dubská, and R. Rudolf Kajan, "Five shades of grey for fast and reliable camera pose estimation.," *CVPR*, 2013.

[10] D. Hoiem, A. A. Efros, and M. Hebert, "Geometric context from a single image.," *ICCV*, 2005.

[11] N. Cornelis, B. Leibe, K. Cornelis, and L. Van Gool, "3D urban scene modeling integrating recognition and reconstruction.," *IJCV*, 2008.

[12] N. Dalai and B. Triggs, "Histograms of oriented gradients for human detection.," *CVPR*, 2005.

[13] R. Martin and O. Arandjelović, "Multiple-object tracking in cluttered and crowded public spaces.," *ISVC*, 2010.

[14] O. Arandjelović, "Crowd detection from still images.," *BMVC*, 2008.

[15] D. Tran and A. Sorokin, "Human activity recognition with metric learning.," *ECCV*, 2008.

[16] Philips Electronics N.V., "A surveillance system with suspicious behaviour detection.," *Patent EP1459272A1*, 2004.

[17] G. Lavee, L. Khan, and B. Thuraisingham, "A framework for a video analysis tool for suspicious event detection.," *MTA*, 2007.

[18] O. Arandjelović, "Contextually learnt detection of unusual motion-based behaviour in crowded public spaces.," *ISCIS*, 2011.

[19] intellvisions, "iQ-Prisons," `http://www.intellvisions.com/`, Accessed March 2015.

[20] O. Arandjelović, D. Pham, and S. Venkatesh, "The adaptable buffer algorithm for high quantile estimation in non-stationary data streams.," *IJCNN*, 2015.

[21] iCetana, "iMotionFocus.," `http://www.icetana.com/`, Accessed March 2015.

[22] O. Arandjelović, D. Pham, and S. Venkatesh, "Stream quantiles via maximal entropy histograms.," *ICONIP*, 2014.

[23] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features.," *CVIU*, 2008.

[24] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors.," *TPAMI*, 2004.

[25] B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision.," *IJCAI*, 1981.

[26] O. Arandjelović, D. Pham, and S. Venkatesh, "CCTV scene perspective distortion estimation from low-level motion features.," *TCSVT*, 2015.